

EP34694 C

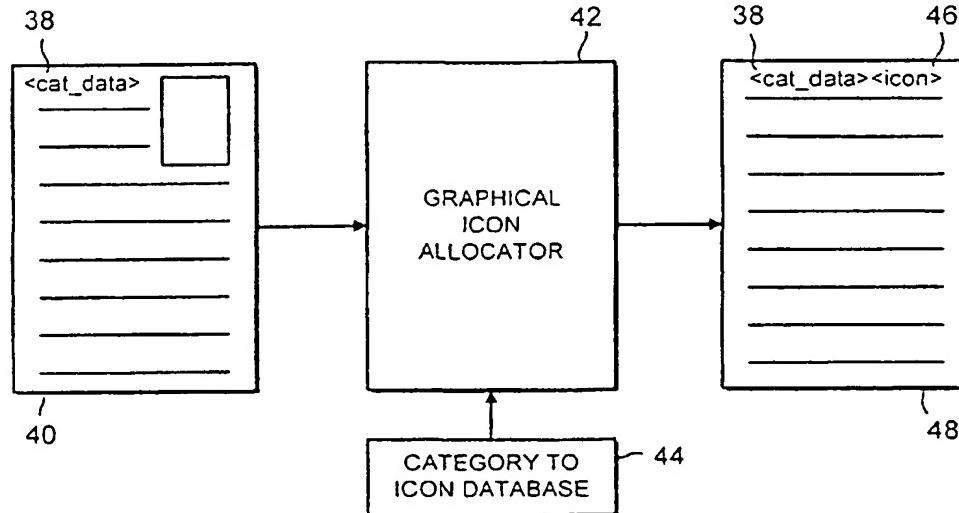
PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : G06F 17/30		A2	(11) International Publication Number: WO 00/68833
			(43) International Publication Date: 16 November 2000 (16.11.00)
(21) International Application Number: PCT/GB00/01535		(81) Designated States: JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(22) International Filing Date: 19 April 2000 (19.04.00)			
(30) Priority Data:		Published	
9910683.3	7 May 1999 (07.05.99)	GB	<i>Without international search report and to be republished upon receipt of that report.</i>
9910684.1	7 May 1999 (07.05.99)	GB	
9910679.1	7 May 1999 (07.05.99)	GB	
9910682.5	7 May 1999 (07.05.99)	GB	
9910685.8	7 May 1999 (07.05.99)	GB	
(71) Applicant (<i>for all designated States except US</i>): ARGO ACTIVE LIMITED [GB/GB]; 7 Dukes Court, Chichester, West Sussex PO19 2FX (GB).			
(72) Inventors; and			
(75) Inventors/Applicants (<i>for US only</i>): JELBERT, Richard [GB/GB]; 37 Bognor Road, Chichester, West Sussex PO19 2NG (GB). TRIBBECK, Jason, Paul [GB/GB]; 36 Nettlecombe Avenue, Southsea, Portsmouth PO4 0QW (GB).			
(74) Agent: ROBINSON, Nigel, Alexander, Julian; D Young & Co., 21 New Fetter Lane, London EC4A 1DA (GB).			

(54) Title: CATEGORISING DATA



(57) Abstract

A data processing system modifies a data file representing a document (14) and including link data items (24) by identifying keywords within the link data items and mapping these key words to category data (38) that is then inserted into the document. The document may be an internet web page and the link data item a hypertext link.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

CATEGORISING DATA

This invention relates to the field of data processing. More particularly, this invention relates to data processing systems for manipulating data files representing documents, such as, for example, internet web pages.

5 The overwhelming majority of information currently available via the internet in HTML and other mark-up languages has been designed for display on a desktop computer monitor having a typical resolution of 640 by 480 or 1024 by 768 pixels. There is a need to make this existing material available for use with devices having a much reduced display resolution, such as 120 by 90 pixels. Such low display 10 resolution devices are typically mobile telephones or personal digital assistance.

One problem that arises when displaying this material on display devices having a much lower resolution than was originally envisaged is that it becomes increasingly difficult to navigate through the document. When using a lower resolution display device, such that a smaller portion of the document may be viewed 15 at any one time compared with that envisioned by the author, the relationship between items on the page and the ability to navigate through the material is less clear.

It is known to provide systems that access web documents and analyse the text material of the document in an effort to categorise that document. These systems are often used by search engines which automatically retrieve internet web pages, analyse 20 and categorise the content of those pages for addition to their index, follow all the links from those pages and then categorise those further pages until there are no more links to follow. This process is sometimes referred to as "spidering". Spidering consumes considerable processing and band width resources and is not well suited to real time operation.

25 Viewed from one aspect the present invention provides a method of adding category data to a data file representing a document, said data file including at least one link data item specifying a linked location within said document or another document, said method comprising the steps of:

- (i) searching said document for a link data item;
- 30 (ii) identifying one or more link keywords within said link data item;

(iii) mapping said one or more link keywords to category data for said link data item; and

(iv) adding said category data in a manner associated with said link data item to said document.

5 The invention recognises that highly useful categorising data that may assist in navigation through a document can be derived from analysing the link data within the document. The link data pointing to another location within that document or another document often contains sufficient keyword information in itself that an accurate categorisation of the target of that link can be achieved without the need to follow the
10 link. This greatly reduces the processing and bandwidth requirements making the system much better suited to real time dynamic operation. In contrast to spidering techniques, a target document is characterised by an analysis of the link data pointing to that document rather than by retrieving and analysing the document itself.

15 It will be appreciated that whilst the principle of the invention can be used with many different sorts of document including link data, the invention is particularly well suited to documents in the form of a mark-up language document where the link data is a hypertext link.

20 Hypertext links usually include a universal resource identifier that is used by the software responsive to that hypertext link to access the linked location. The universal resource identifiers are assigned by the author of the document and are usually not displayed as they have a syntax and form not intended to be read by a human. Nevertheless, this invention exploits the realisation that a universal resource identifier is often specified by the author in a manner that contains a good source of keyword data useful in accurately categorising the target for that link. Accordingly, in
25 preferred embodiments of the invention the hypertext link includes a universal resource identifier and said step of identifying includes identifying link keywords within said universal resource identifier.

A hypertext link may also include associated text that is intended for display within the document at the point at which the hypertext link is inserted. This text for
30 display is also a good source for keyword information useful in categorising the target of the link.

In order to increase the efficiency of the mapping step, it is preferred that said category data is associated with a category data entry within a catalogue of category data.

The catalogue of category data may be efficiently organised such that each 5 category data entry is associated with one or more category keywords and said step of mapping includes matching said link keywords with category keywords.

The selection of a particular category from among several possible categories that may be associated with keywords is assisted in preferred embodiments in which said step of matching includes calculating a score representing a degree to which said 10 item keywords match a category data entry.

The accuracy of the category identification can be improved in embodiments in which each category keyword has an associated score value.

Associating a score with each category keyword allows category keywords that are strongly correlated with a particular category to be distinguished from words only 15 weakly correlated with that category and this difference exploited in accurately identifying the correct category.

In some cases the keyword data may be sufficient to identify a highly specific category whilst in others only a much less specific category may be identified with any confidence. This is accommodated by providing the catalogue in the form of a 20 hierarchical catalogue. In this way, if highly specific characterising data is available from the keywords, then a specific category low down in the hierarchy can be identified, whereas if less specific data is identified, then only a high level category may be identified. Providing the categorising data in a hierarchical form also allows an improved degree of flexibility in the way that this added category data may 25 subsequently be utilised by other systems. For example some other systems may wish to exploit the greatest level of detail available within the category data whilst others more usefully utilise only higher level, less specific divisions between categories.

In some circumstances the link data may not provide enough useful information for categorising the target of the link. In order to deal with this, preferred 30 embodiments of the system are such that said step of identifying includes detecting if said link keywords exceed a minimum level of specificity in identifying associated

category data and, if not, accessing title data for said linked location and adding keywords from said title data to said link keywords for use in said step of mapping.

Applying a set of rules to determine whether the minimum level of specificity is met and then, if necessary, accessing title data from the linked location increases the likelihood of a proper categorisation being achieved without incurring the processing and bandwidth overhead of fetching and analysing the whole of the content of the target location.

The detection of the minimum level is preferably in the form of the application of a sequence of tests applied to the keywords.

10 In preferred embodiments of the invention said step of identifying includes pre-processing said link data item to increase a level of specificity of words within said link data item in identifying associated category data.

Such pre-processing can significantly improve the usefulness of the link data item, such as by stripping out punctuation, computer language prefixes and/or suffixes, 15 non-specific small words (e.g. "a", "the", etc) and the like.

The category data added to the document is preferably added in the form of a metatag. A metatag can be found and processed by subsequent systems and yet does not alter the display appearance of the document in itself.

Whilst the invention is useful in many different circumstances, it is particularly 20 useful when the document is an internet web page and the data file is a HTML data file.

The categorisation of the invention may conveniently be performed by a proxy server disposed within a computer network between the source of the document and the client for that document. The proxy server typically will have a high processing 25 capacity and so be able to deal with the processing load of the categorisation better than a client computer.

In contrast, where the processing capabilities of the client computer are greater, there are advantages in performing the categorisation locally on the client computer as the system is then self-contained with the client computer being responsible for all of 30 the desired augmentation of the source document.

The category data added to the document can be used for various different purposes, such as filtering on the basis of content or modifying other content (e.g.

content specific advertising), but is particularly useful for assisting in the modification of documents for display on a user display with different capabilities than that for which the document was originally intended or when the document is display independent. Such different user displays are common on wireless mobile devices
5 when the original documents were intended for display on a full computer monitor.

Viewed from another aspect the present invention provides apparatus for adding category data to a data file representing a document, said data file including at least one link data item specifying a linked location within said document or another document, said apparatus comprising processing logic performing the steps of:

- 10 (i) searching said document for a link data item;
- (ii) identifying one or more link keywords within said link data item;
- (iii) mapping said one or more link keywords to category data for said link data item; and
- (iv) adding said category data in a manner associated with said link data item to
15 said document.

Viewed from a further aspect the invention also provides a computer program storage media storing a computer program for controlling a data processing apparatus to perform the categorisation technique in accordance with the above.

An embodiment of the invention will now be described, by way of example
20 only, with reference to the accompanying drawings in which:

- Figure 1 schematically illustrates a computer network;
- Figure 2 schematically illustrates a system for adding categorising data to a data file representing a document;
- Figure 3 illustrates a link data item and associated keywords;
- 25 Figure 4 schematically illustrates a hierarchical category database;
- Figure 5 illustrates a category data entry;
- Figure 6 illustrates how a web page may be modified using category data to filter out links known to be unwanted or less wanted by a user;
- Figure 7 is a flow diagram illustrating the addition of category data to a
30 document;
- Figure 8 schematically illustrates a system for adding output graphical data to a document;

Figure 9 illustrates a low resolution display device showing a document before and after addition of icons in accordance with category data;

Figure 10 is a flow diagram illustrating the addition of output graphical data items in association with link data within a document;

5 Figure 11 schematically illustrates modifying display text associated with a link data item into a more readable form;

Figure 12 shows a flow diagram illustrating the process of modifying display text into a more readable form;

10 Figure 13 illustrates various examples of text modifications that may be performed;

Figure 14 illustrates an unmodified hierarchy of documents including repeated components;

Figure 15 illustrates a modified form of the hierarchy of Figure 14 in which repeated components have been removed;

15 Figure 16 illustrates the comparison between a universal resource identifier based hierarchy and a session based hierarchy;

Figure 17 is a flow diagram showing the process for removing repeated components within a hierarchy; and

20 Figure 18 schematically illustrates a data processing apparatus that may serve as a client computer.

Figure 1 illustrates a computer network 2. This computer network 2 may be a portion of the internet in which internet web pages in the form of HTML data files are transmitted between source servers 4 and client computers 6, 8. A proxy server 10 is disposed between the source servers 4 and the client computers 6, 8. The client 25 computer may be a normal desktop computer 6 for which the internet web pages are primarily designed and intended. The client computer may also be in the form of an internet-enabled mobile telephone 8 connected via a radio link 12 to the computer network 2.

30 The mobile phone 8 connects via the proxy server 10, and the proxy server 10 may detect (e.g. via user id and password details) that the link from the mobile phone 8 as a client computer is to a device having a smaller and less capable display than a full desktop computer 6. Accordingly, the proxy server 10 is able to perform additional

processing steps on the internet web pages fetched from the source servers 4 before they are passed to the mobile telephone 8 so that they can be adapted to be more usefully displayed on the mobile telephone 8. It will be appreciated that if the processing capabilities of the mobile telephone 8 were greater and the radio bandwidth 5 sufficient, then the full internet web pages could be transmitted to the mobile telephone 8, which may then conduct its own processing of those pages to put them into a form more suitable for display on its smaller display output.

Figure 2 schematically illustrates how a data file representing a source document 14 may be processed by a link categoriser 16 to generate an output 10 document 18 that has category data added to it. It will be appreciated that the link categoriser 16 will typically take the form of a general purpose computer executing software written to perform the function of adding the category data to the documents. The link categoriser 16 uses a category-to-keyword database 20 which enables keywords identified within the source document 14 to be mapped to appropriate 15 categories. The category-to-keyword database 20 can be in the form of a hierarchical database with each category data entry having the keywords associated with that category data entry related thereto and with score values for each associated keyword. The link categoriser 16 also uses a user-to-category database 22 which enables the link categoriser to perform other functions, such as modifying the source document in a 20 way that removes or adds data known to be of particular interest the user concerned.

Figure 3 illustrates a link data item 24 that is typically embedded within a HTML document. The link data item 24 includes a universal resource identifier 26 and display text 28. If display text 28 is present, then this is what will be displayed as the hypertext link in the document. If display text 28 is not present, then the universal 25 resource identifier 26 will be displayed.

The keywords within the link data item 24 are identified by processing the link data item 24 by removing all punctuation and replacing this with spaces. The resulting stream of keywords 30 can then be input to the keyword-to-category matching database 20. The category-to-keyword database 20 can be arranged as a relational 30 database making the analysis of the keywords sufficiently rapid to be performed in real time by the proxy server 10.

Figure 4 schematically illustrates the hierarchical nature of the category database 20. In particular, a category such as "Transport" can be broken down into a number of sub-categories such as "Car", "Motorcycle", "Bicycle", "Lorry", and "Van". Each of these sub-categories can be further broken down as illustrated. The hierarchy could have a varying depth depending upon the required degree of specificity traded off against the processing and data storage requirements as well as the likelihood of a highly specific categorisation in fact being correct.

Figure 5 schematically illustrates a particular category data entry within the category-to-keyword database 20. In this case, the category data 32 is associated with a sequence of keywords 34 each having an associated score value 36. The keywords 30 with the link data item 24 are matched against the keywords 34 and the score values 36 for each match of a category data entry 32 added together. The category data entry 32 having the highest score is deemed to be the match.

Returning to Figure 2, when the category data entry 32 that produces the best match has been identified, then category data 38 in the form of a metatag is inserted into the document 18 in association with the link data item 24 that has been analysed. The category data 18 thus gives a representation of the subject matter to which the link data item 24 relates. This information is highly useful to other processes performed by the proxy server 10. In particular, the proxy server 10 might automatically insert a graphical item before each hypertext link to assist in faster recognition of links of interest. The proxy server 10 could filter out categories that are known to be unsuitable or undesired for the user, for example if the reader is known within the user-to-category database 22 to not want information concerning cars. The proxy server 10 can also record information regarding the categories of links followed by a user while viewing hypertext documents and so assemble a profile of the user's interest such that other material of possible interest to the user, such as targeted advertising, may be presented to the user. Another use that can be made of such user profiling information is pre-fetching of information relevant to the user's interests. Using pre-fetching, the proxy server 10 may automatically collect and store information that the user is likely to want to view before they request it. If they do then request this information, it can be delivered more quickly. If they do not request the information, then the information can be discarded.

Figure 6 shows how an original web page 80 containing ten hypertext links can be modified into a page 82 more suited to display using a smaller display window 84 by the removal of hypertext links detected as either not wanted or less likely to be wanted by a user. This is done by comparing the category data 38 associated with 5 each link with the user preference data stored in the user to category database 22. The user to category database 22 can contain preference data obtained by the user specifying categories of link in which they are not interested and do not wish to display. Alternatively or additionally, the user to category database 22 can be automatically built up by the proxy server 10 keeping a record of the categories of the 10 links that a user follows, e.g. by dynamically user profiling the categories of interest. Thus, categories stated or observed to be of little interest to a user can be removed from the page 82 so making better use of the limited bandwidth and display resources. This sort of content filtering may also be used to block material, such as by a parent wishing to prevent access to unsuitable material by a child.

15 Figure 7 is a flow diagram illustrating the process of adding category data to a source document. At step 52, the source document is fetched via the network link from the source server 4. The proxy server 10 at step 54 processes the source document to identify the link data items 24 within it and isolate the keyword data within those link data items 24. At steps 56 and 58, the proxy server applies a series of 20 rules to the keywords identified within the link data item 24 to determine whether they are sufficiently specific to enable a proper categorisation to be made. An example of the rules applied are as follows:

- 1) Initially everything is neat, i.e. is initialized in a state termed "neat";
- 2) It is ruled as being not neat if the length of the text is greater than 10
25 AND the length to space ratio is greater than 10:1;
- 3) It is ruled as being neat if the text is "entertainment";
- 4) It is ruled as being not neat if the text is "image" followed by a
number;
- 5) It is ruled as being not neat if the length of the text is less than 4
30 characters;
- 6) It is ruled as being not neat if the number of underscores exceeds the
number of spaces;

- 7) It is ruled as being not neat if the text begins with "http://";
- 8) It is ruled as being not neat if the text is enclosed with quotes;
- 9) It is ruled as being not neat if the text begins with "image map";
- 10) It is ruled as being not neat if the text is "default".

5 In addition, there are additional rules that may be added for specific geographical locations, e.g:

- 11) It is ruled as neat if the text contains "Island";
- 12) It is ruled as neat if the text contains "Kanagawa-Ken".

10 Both of these (and also some of the specific rules) may be added in a category such as 'rules specific to sites'.

If sufficient information is present, then processing proceeds to step 60. If sufficient information is not present, then the proxy server 10 fetches the title data of the target location identified by the link data item 24 to derive additional keywords from that title data. The entire document indicated by the link data item need not be fetched. This contrasts to spidering in which the entire document pointed to by a link data item is fetched and analysed.

20 At step 60, the proxy server/link categoriser 16 looks up the keywords identified within the category-to-keyword database 20 and scores each possible category. At step 62, the category with the highest score is selected to be associated with the link data item 24. At step 64, a metadata tag identifying the category selected at step 62 is inserted into the document in association with the link data item 24.

Figure 8 schematically illustrates a system for modifying the graphical data contents of a document. A source document 40 is accessed from a source server 4 via an internet link. The source document 40 is in the form of a HTML document 25 representing an internet web page. The source document 40 may contain GIF files, JPEG files and bitmap files as part of its source graphical data content. The source document 40 includes category data 38 classifying the link data items 24 as added by the processing discussed above.

A graphical icon allocator 42 receives the source document 40 and removes all 30 or some of the source graphical data items. The graphical icon allocator 42 then accesses a category-to-icon database 44 where icons suitable for association with each link data item 24 within the source document 40 are identified using the category data

38 embedded within the source document 40. When an output graphical data item has been identified from the category-to-icon database 44, then data identifying this icon 46 is inserted as a metatag into the output document 48. The data identifying the output graphical data item 46 may be merely an identifier for an icon which is built 5 into the known display device 8, or alternatively it may be data giving sufficient information to specify the appearance of the icon without this already being embedded within the display device 8.

It will be appreciated that the graphical icon allocator 42 will typically take the form of software operating on a general purpose computer, such as the proxy server 10. If the processing capabilities of the client computer 8 are sufficient and sufficient bandwidth is available, then the source document 40 may be transmitted to the client computer 8 in its entirety and the processing illustrated in Figure 6 performed wholly 10 within the client computer 8.

Figure 9 illustrates a small low resolution display device 50, such as the small 15 LCD display of a mobile telephone 8. The left hand portion of Figure 7 illustrates a text-only web page showing a series of hypertext links with all of the graphical data from the source page removed. The usability of such a display is poor compared to the original source document 40 as users derive considerable information from the graphical data content of a page.

Using the present invention, the links within the page can be categorised and 20 then appropriate icons associated with each link. These icons can be built into the mobile telephone 8 itself such that they do not need to be transmitted to the client computer in their entirety. A code identifying a particular built-in icon can merely be added as the data 46 in the output document 48.

Figure 10 is a flow diagram illustrating the processing of graphical data items. At step 66, the proxy server 10 fetches a source document 40. At step 68, the proxy server/graphical icon allocator 42 removes all non-text data from the source document 40. At step 70, the graphical icon allocator maps the category data 38 to icons to be associated with the link data item 24 using the category-to-icon database 44. At step 30 72, the icon identifying data is inserted as a metatag 46 within the output document 48. At step 74, the resulting output document 48 including text data and associated icon data is transmitted to the client computer 8. At step 76, the client computer 8

processes the received document and displays the text with its associated icons next to the link data items. The icons can be built-in icons within the client computer 8 itself.

Figure 11 illustrates a source document 78 in the form of an internet web page intended by the author to be displayed and manipulated using a conventional personal computer. Within the document 78 there is a link data item 80 in the form of a hypertext link to a large image file. A small thumbnail representation 82 of the full image file is also shown. When a user accesses this web page 78 on a conventional personal computer, then the thumbnail representation 82 in combination with the display text of the link 80 gives sufficient information for the user to understand the link being made. However, if the web page 78 is modified to produce a modified page 84 in which graphical data has been removed, then the initial display text 86 associated with the link 80 may not be sufficient to enable a user to properly understand the connection being made.

The system identifies the links within the web page 78 and performs tests upon the initial display text associated with each link to determine characteristics indicative of insufficient readability. In the case of the initial display text 86 shown in Figure 11, then this may fail the test of comprising too many characters within a word or of including a capital letter following a lower case letter within the middle of a word. The initial display text 86 having been identified as not sufficiently readable, the title 88 of the page to which the link relates is accessed and this title used as further text in place of the initial display text 86. The title 88 is itself subject to an assessment of its readability and only if it passes this determination does it remain as a replacement for the initial display text 86. If the further text 88 fails the readability test, then the initial display text is reverted to for the link 80.

The above technique uses a system of computer software through which users are required to fetch hypertext documents that they wish to read. Typically this is in the form of an intermediate "proxy server", but a stand-alone mode of operation can also be envisaged. The system processes the hypertext pages as they are transferred from the storage location to the reader. After identifying the links in the hypertext document, the textual part of the hypertext link (i.e. the text which the user would select in order to go to the linked document) is checked to see if it is readable. This can be done in a number of ways, including (but not limited to):

- the number of underscores is greater than the number of spaces;
- the text is less than a certain number of characters long;
- the text is longer than a certain number of characters long;
- the average number of characters per word is greater than a certain limit;
- 5 • the text contains words which have capital letters after lowercase letters in the same word (e.g. gooSE);
- the text contains words which are not in a dictionary;

A combination of the above rules can be used to score the link in terms of readability, and if the score is above a threshold, then an alternative to the text is sought. This can also be done in several ways, including (but not limited to):

- fetching the linked hypertext document and retrieving the document's title (should one exist), or the first line of the text in the document;
- substituting the text with different text from a dictionary (stored in a file coupled to the proxy server, e.g. a keyword to further text mapping);
- 15 • replacing with the title of the current document (should one exist);
- using a filename with its file type suffix removed.

If the further text that is to replace the initial display text is deemed more unreadable than the initial display text, then the initial display text is kept in place, and either no substitution takes place, or an alternative substitution is used.

20 Figure 12 shows a flow diagram illustrating the technique of improving the readability of the display text associated with links.

At step 90 a page to be accessed is fetched from a remote computer server. At step 92 the fetched page is searched to detect link data items (hypertext links) and the initial display text associated with these links is determined. At step 94 the readability 25 rules described above are applied to the initial display text of each link. At step 96 a determination is made as to whether or not the initial displayed text passes the readability rules. If the initial display text does pass the readability rules, then the process proceeds to step 98 where the output page is generated.

If the initial display text does not pass the readability rules at step 96, then step 30 100 is used to replace the text with further text derived in dependence upon the link item data, such as by using the replacements described above. These candidate

replacements can be applied in turn with each candidate replacement being tested by steps 102 and 104 to determine whether or not it passes the readability test. If it does pass the readability test at step 104, then the replacement candidate is used as the further text to replace the initial display text within the link data item and an output page including this further text is produced at step 98. If the candidate replacement text does not pass the readability test, then the next candidate replacement text will be tried providing step 106 does not determine that all the candidates have been exhausted. If step 106 does determine that all the candidate replacement text have been exhausted, then step 108 reverts to the initial display text and the output page is produced using this initial display text at step 98.

Figure 13 schematically illustrates how some initial display text may be modified into forms more readily readable. In example A, a file name containing a mixture of numbers and underscore characters and exceeding a predetermined length is replaced by the title of the page to which it points. In example B, an initial display text that is too short to be useful is replaced with category data associated with the link and derived as described above. In example C, an initial display text that is too long to be usefully displayed on a mobile telephone is replaced by a text that uses keywords selected from the initial longer text. Finally, in example D, a file name is replaced by the file name minus its file type suffix.

As previously described, it will be appreciated that the processing described above to improve the readability of the display text associated with a link data item may be performed either on a proxy server using the superior processing and storage capabilities of that proxy server, or upon the client device itself. As the client devices improve in their capability, it will be natural for more processing to take place upon the client device and so remove the need for the connection to have to be made through a particular proxy server.

Figure 14 schematically illustrates an internet web site in the form of a hierarchy of documents. Each page has an associated universal resource identifier 110 with a form similar to a directory/subdirectory structure. The hierarchy illustrated starts with a company home page 112 and progresses to a products page 114 and a support page 116 via respective hypertext links 118 and 120. The hypertext links 118 and 120 together with a home page link 122 form a navigation bar that appears on all

of the pages of the web site. A company logo 124 and a standard footer text 126 also appear on all pages of the web site.

The product page 114 includes two further hypertext links 128 and 130 that respectively point to pages 132 and 134 giving details of retail and wholesale products.

- 5 Each of the pages 112, 114, 116, 132 and 134 also includes its own unique text.

It will be appreciated that when processing and bandwidth resources as well as display device resources are limited, then the repeated transmission, processing and display of items such as the company logo 124 and the footer text 126 represent a significant overhead. Assuming that a user enters the site at page 112, then they are 10 initially presented with the opportunity to progress to the support page. If instead the user progresses to the products page 114, then it is reasonable to assume that they are not interested in support. Accordingly, it is wasteful to display the link 120 to the support page 116 on the product page 114 as well as on the home page 112.

Figure 15 illustrates the web site shown in Figure 14 but this time modified 15 such that repeated components lower down in the hierarchy are removed, i.e. in this arrangement components appear upon their first occurrence when moving down the hierarchy but are thereafter removed. As an example, the company logo 124 appears on the home page 112, but does not appear on any of the pages lower in the hierarchy.

Similarly the footer text 126 appears only on the home page 112 and has been removed

- 20 from the lower pages. The links 118, 120 and 122 that form the navigation bar appear only on the home page 112. On the lower pages, a link 136 is added linking to the top page in the hierarchy. Where there is a page above the current page that is not the top page, then an uplink 138 is also added.

It will be seen from Figure 15 that the content of the pages below the home 25 page 112 has been significantly reduced so enabling them to be more rapidly transmitted to a client computer and conveniently and rapidly manipulated on that client computer. Nevertheless, all of the content of the original web site illustrated in Figure 14 is present within the modified web site shown on Figure 15 at some point within that web site.

30 Figure 16 schematically illustrates how a web site may be placed into a hierarchy based upon the universal resource indicators as compared to a session hierarchy. On the left hand side of Figure 16 is shown a hierarchy derived from the

universal resource identifiers. The letters next to each node indicate a unique page. The vertical position within the illustrated hierarchy denotes the position within the hierarchy. The numbers next to each node represent the order in which the pages are accessed during a user session. With the hierarchy based upon the universal resource identifier, page a is at the top of the hierarchy and page e is towards the centre. Compared to the universal resource identifier hierarchy, the session hierarchy illustrated in the right hand portion of Figure 16 shows a hierarchy in which the first pages to be accessed are disposed higher within the hierarchy. Accordingly, since the first page accessed (e.g. through a bookmark) was page e, this is at the top of the hierarchy. A user may subsequently traverse the entire web site in the order shown by the numbers. The pages are arranged in the session hierarchy according to these numbers with pages at the same horizontal level indicating the same position within the hierarchy.

Hypertext documents are viewed in some sequence by each reader, moving from one to another by choosing "links" within each page. Where some information is presented on an early page and then ignored by the reader, it is reasonable to assume that they are not interested in it. Also, many modern hypertext document systems (sometimes called "web sites") are designed in a hierarchical form. There may be pages to list the sections of the web site, and more to list each sub-section, followed by pages containing actual content. Either such a hierarchy or the historical tracking of a user's reading can be employed to assist the system predicting which pages a reader should already have read, if historical tracking information has not been recorded for them.

The present technique uses a system of computer software, through which users are required to fetch hypertext documents that they wish to read. Typically this is in the form of an intermediate "proxy server", but a stand-alone mode of operation can also be envisaged. The system processes the hypertext pages as they are transferred from the storage location to the reader, removing parts, recording what it has found, and performing other tasks.

Once a hypertext document has been requested by the user and subsequently reviewed by the system, the system examines the hierarchy in which the page exists on the basis of the document's Uniform Resource Identifier (URI). This URI, or some

similar information appropriate to the hypertext system being used, should uniquely identify the page and provide some information about the hierarchy in which it exists. The system fetches each page that is above the requested one in the hierarchy (sometimes called "parent" pages), and makes a note of discrete units of information 5 on each page. It may only note links to other pages, but divisions of other information such as images and/or footnotes are also envisaged. If the reader's activity is being recorded, then pages they have already viewed may be considered instead of parent pages of the current document.

Once a note has been made of the information units on each page, those units 10 that are present on parent pages are removed from the one requested by the reader. One or more new links are added to the current page to ensure that the reader has the opportunity to return to pages which do contain the links, should they wish to use them.

The advantage of this a procedure is that each document will be reduced to a 15 more manageable size without removing significant information from it, and without requiring special preparation by the hypertext author. This is important for small devices that are technically limited and very different from the majority of readers for whom such authors write.

If the system is configured to work with a historical record of pages viewed by 20 the reader, the oldest page considered as part of the link removal may either be the first page seen, the first seen within a certain time, e.g. ten minutes, or the N'th last page, perhaps the tenth last. It would not consider any page viewed after the first viewed of the current page (nor of course would it treat the current page as a previous one). This ensures that if the user goes "back" to a previous page, they will not lose all of the 25 links on it.

Figure 17 is a flow diagram illustrating the above process. At step 140 a target document is accessed. At step 142 the components making up that target document are compared with components known to be in document higher in the hierarchy than the target document. The contents of the components higher in the hierarchy may be 30 determined by fetching those pages in dependence upon their universal resource identifier if they have not already been so fetched or may be determined on a user session basis as previously described.

At step 144 items within the target document found to be repeated components that are present in documents higher in the hierarchy are removed. At step 146 hypertext links to the top of the hierarchy and possibly also to one step up in the hierarchy are added. At step 148 the output page is generated.

5 Figure 18 schematically illustrates a client data processing apparatus, such as a mobile telephone. The client device 150 will typically include a central processing unit 152, a read only memory 154, a random access memory 156, a display driver 158, a display 160, a communications interface 160 and an antenna 162. The central processing unit 152, the read only memory 154, the random access memory 156, the
10 display driver 158 and the communications interface 160 are connected via a common bus 164. The read only memory 154 may form a computer program storage device holding a computer program for controlling the central processing unit 152 to carry out the processing described above where the processing is client based. The random access memory 156 will be used as working storage. The display 160 may be of a
15 reduced size and resolution compared to a typical personal computer, e.g. it may be a low resolution LCD screen as typically found on present day mobile telephones, or just a small display per se. The communications interface 160 illustrated is a wireless interface that is linked to the proxy server 10 via the antenna 162.

CLAIMS

1. A method of adding category data to a data file representing a document, said data file including at least one link data item specifying a linked location within said document or another document, said method comprising the steps of:
 - (i) searching said document for a link data item;
 - (ii) identifying one or more link keywords within said link data item;
 - (iii) mapping said one or more link keywords to category data for said link data item; and
 - 10 (iv) adding said category data in a manner associated with said link data item to said document.
- 15 2. A method as claimed in claim 1, wherein said data file represents said document as a mark-up language document.
3. A method as claimed in any one of claims 1 and 2, wherein said link data items is a hypertext link.
- 20 4. A method as claimed in claim 3, wherein said hypertext link includes a universal resource identifier and said step of identifying includes identifying link keywords within said universal resource identifier.
- 25 5. A method as claimed in any one of claims 3 and 4, wherein said hypertext link includes associated text for display and said step of identifying includes identifying link keywords within said associated text for display.
6. A method as claimed in any one of the preceding claims, wherein said category data is associated with a category data entry within a catalogue of category data.
- 30 7. A method as claimed in claim 6, wherein each category data entry is associated with one or more category keywords and said step of mapping includes matching said link keywords with category keywords.

8. A method as claimed in claim 7, wherein said step of matching includes calculating a score representing a degree to which said item keywords match a category data entry.

5

9. A method as claimed in claim 8, wherein each category keyword has an associated score value.

10. A method as claimed in any one of claims 6 to 9, wherein said catalogue is a hierarchical catalogue.

11. A method as claimed in any one of the preceding claims, wherein said step of identifying includes detecting if said link keywords exceed a minimum level of specificity in identifying associated category data and, if not, accessing title data for 15 said linked location and adding keywords from said title data to said link keywords for use in said step of mapping.

12. A method as claimed in claim 11, wherein said step of detecting includes applying a sequence of tests to said item keywords to detect if said minimum level is 20 exceeded.

13. A method as claimed in any one of the preceding claims, wherein said step of identifying includes pre-processing said link data item to increase a level of specificity of words within said link data item in identifying associated category data.

25

14. A method as claimed in any one of the preceding claims, wherein said category data is added as a metatag.

30 15. A method as claimed in any one of the preceding claims, where said document is an internet web page.

16. A method as claimed in any one of the preceding claims, wherein said data file is a HTML data file.

17. A method as claimed in any one of the preceding claims, wherein said data file
5 is retrieved from a source computer server via a computer network.

18. A method as claimed in claim 17 wherein said steps of searching, identifying,
mapping and adding are performed by a proxy server disposed within said computer
network between said source computer server and a client computer requesting said
10 data file.

19. A method as claimed in claim 17. wherein said steps of searching, identifying,
mapping and adding are performed by a client computer which requests said data file
from said source computer server.

15

20. A method as claimed in any one of claims 18 and 19, wherein said client
computer has a user display with different display capabilities than those of a display
for which said document is intended or said document is display independent.

20 21. A method as claimed in claim 20, wherein said client computer is a wireless
mobile device.

22. Apparatus for adding category data to a data file representing a document, said
data file including at least one link data item specifying a linked location within said
25 document or another document, said apparatus comprising processing logic
performing the steps of:

searching said document for a link data item;
identifying one or more link keywords within said link data item;
mapping said one or more link keywords to category data for said link data
30 item; and
adding said category data in a manner associated with said link data item to
said document.

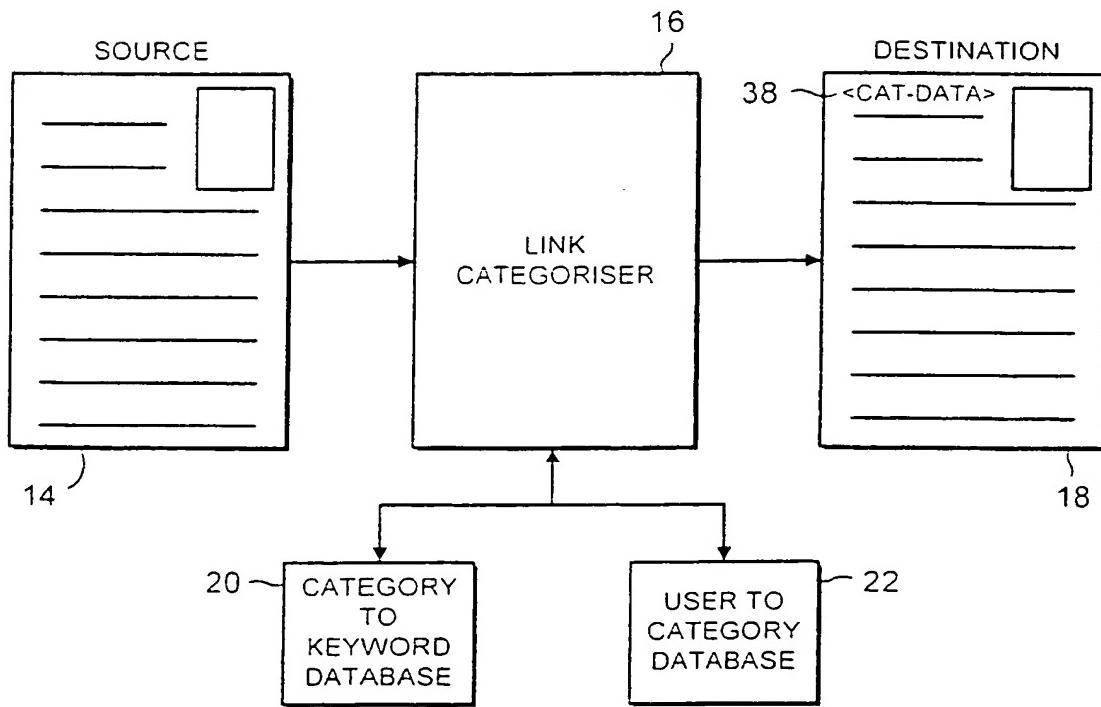
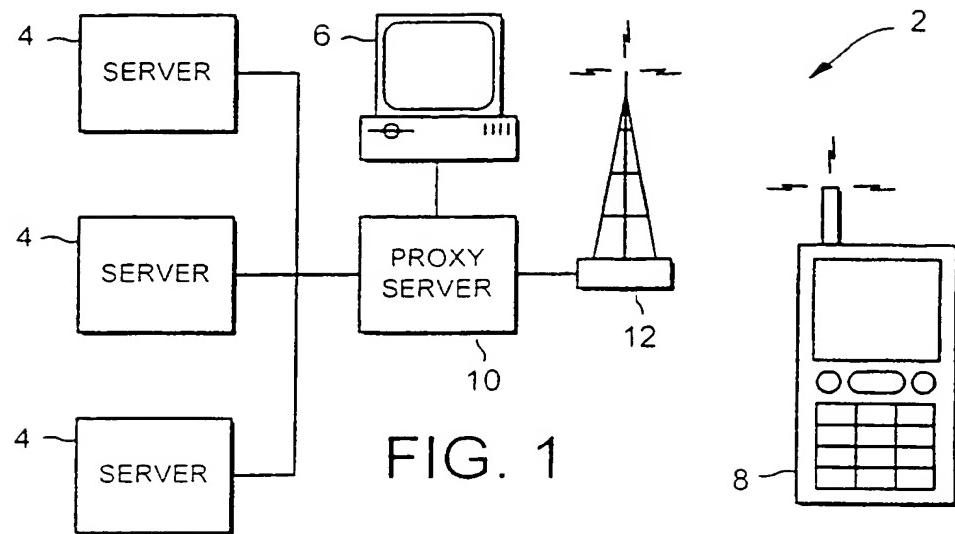
23. Apparatus as claimed in claim 22, wherein said data file is retrieved from a source computer server via a computer network.

5 24. Apparatus as claimed in claim 23, wherein said processing logic is part of a proxy server disposed within said computer network between said source computer server and a client computer requesting said data file.

10 25. Apparatus as claimed in claim 23, wherein said processing logic is part of a client computer which requests said data file from said source computer server.

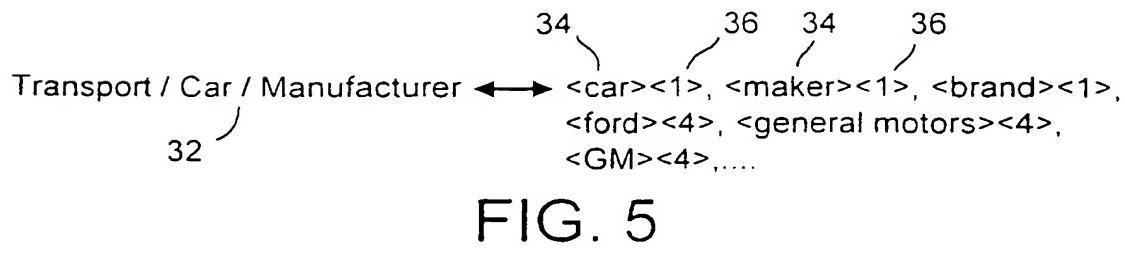
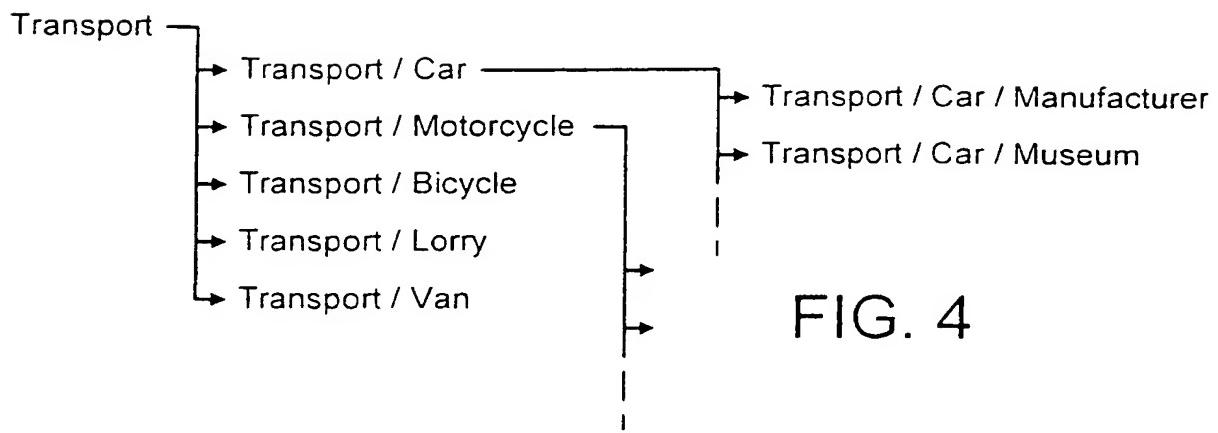
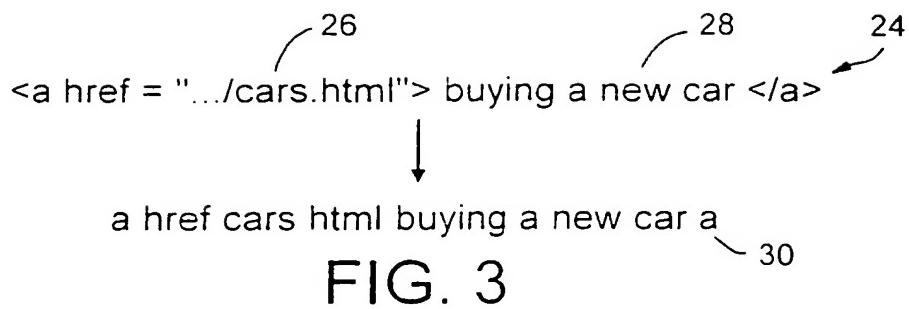
26. A computer program storage medium storing a computer program for controlling a data processing apparatus to perform the method as claimed in any one of claims 1 to 21.

1 / 9



SUBSTITUTE SHEET (RULE 26)

2 / 9



3 / 9

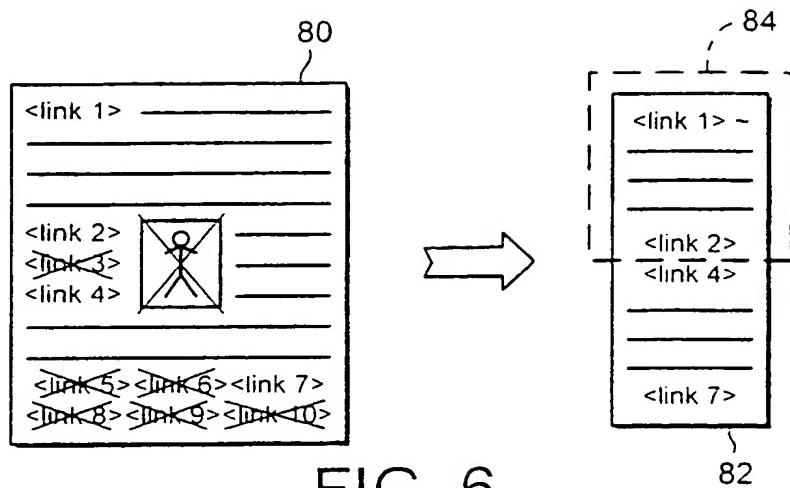


FIG. 6

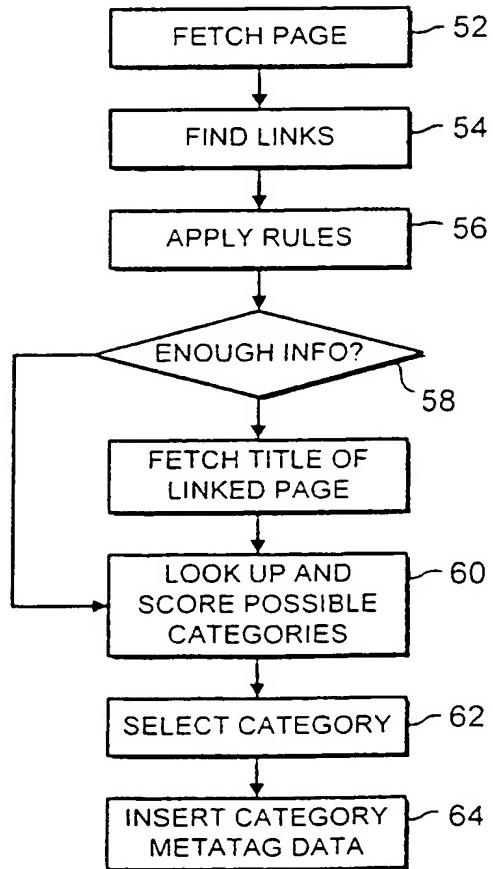


FIG. 7

SUBSTITUTE SHEET (RULE 26)

4 / 9

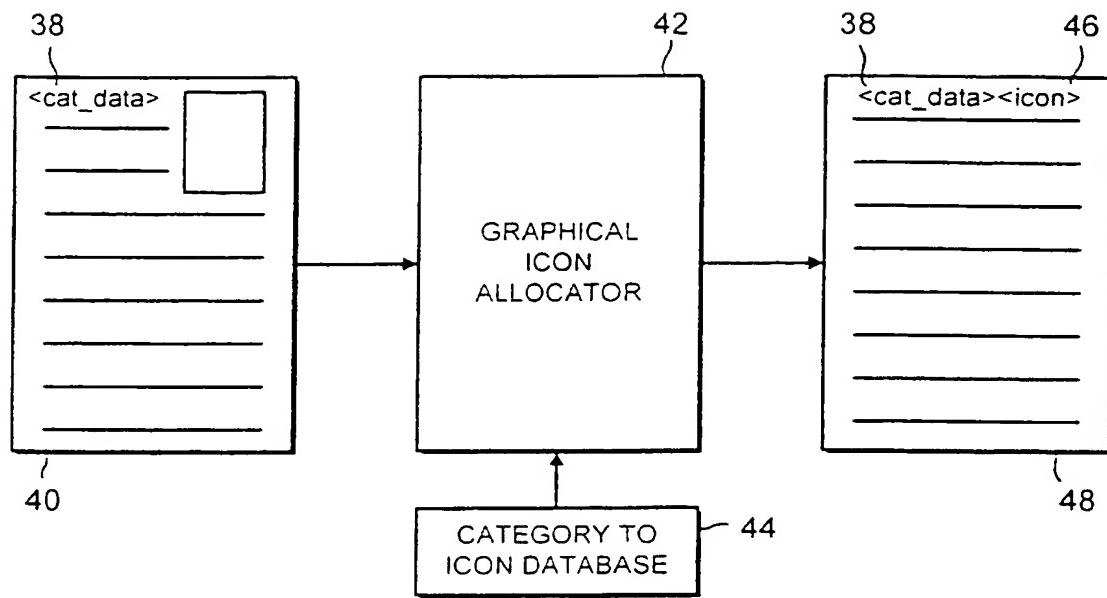


FIG. 8

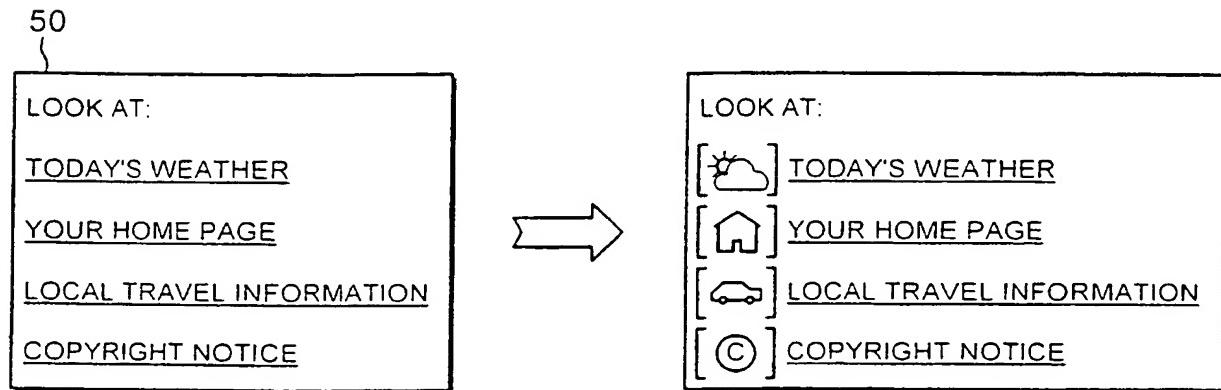


FIG. 9

SUBSTITUTE SHEET (RULE 26)

5 / 9

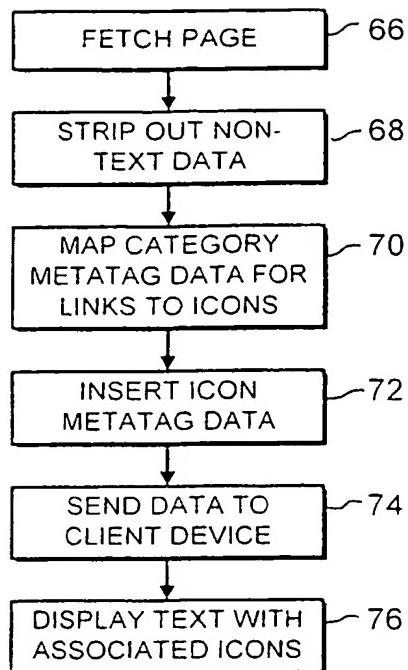


FIG. 10

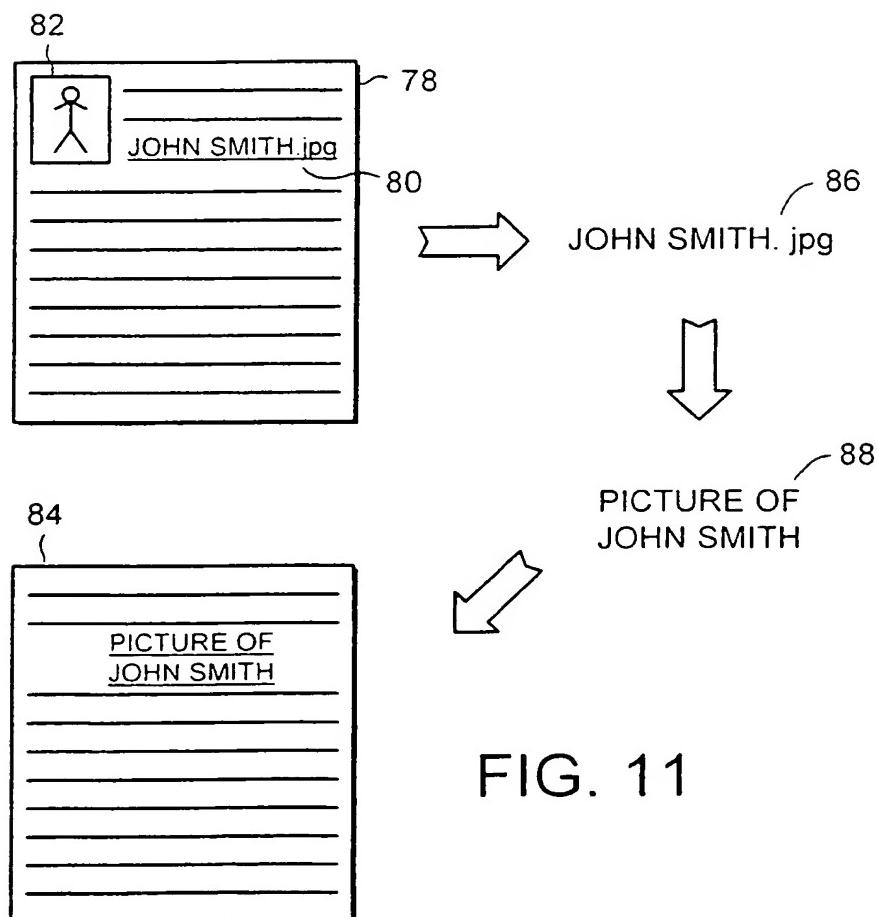


FIG. 11

SUBSTITUTE SHEET (RULE 26)

6 / 9

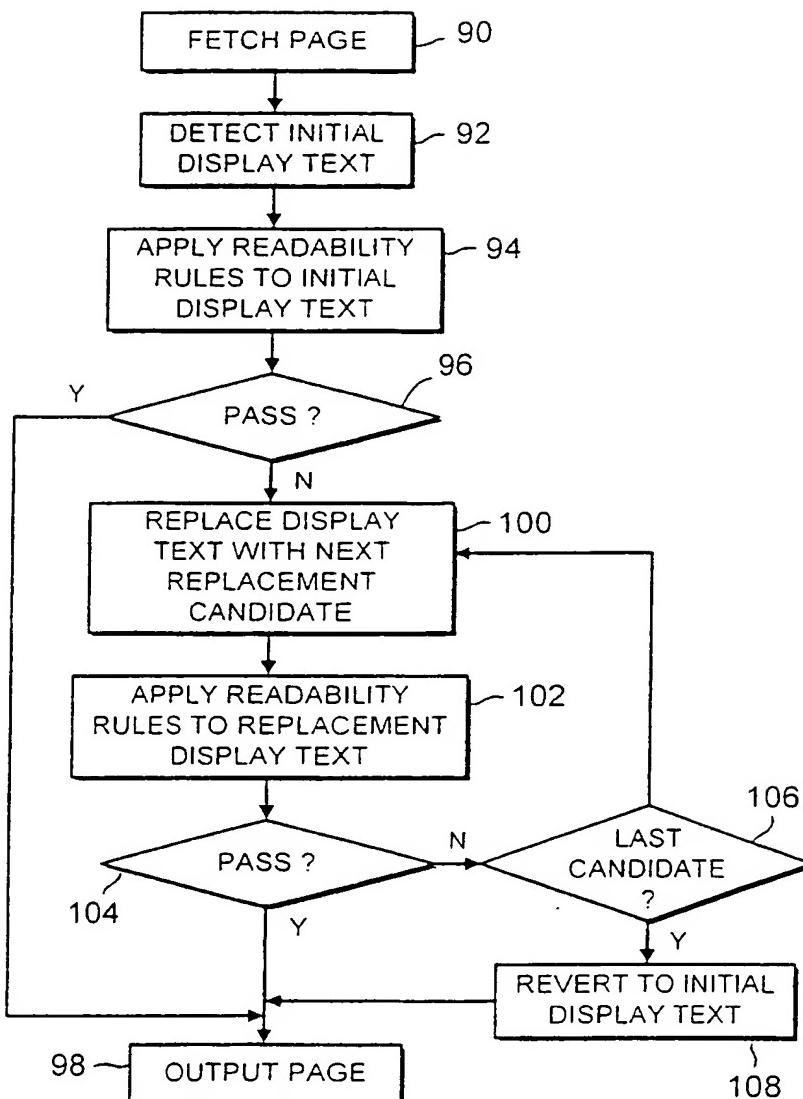


FIG. 12

- A "file_001234.doc" → "contact details page"
- B "link 1" → "transport - cars"
- C "please select this link to see
a full list of our products
available in your country" → "select products country"
- D "products.doc" → "products"

FIG. 13

SUBSTITUTE SHEET (RULE 26)

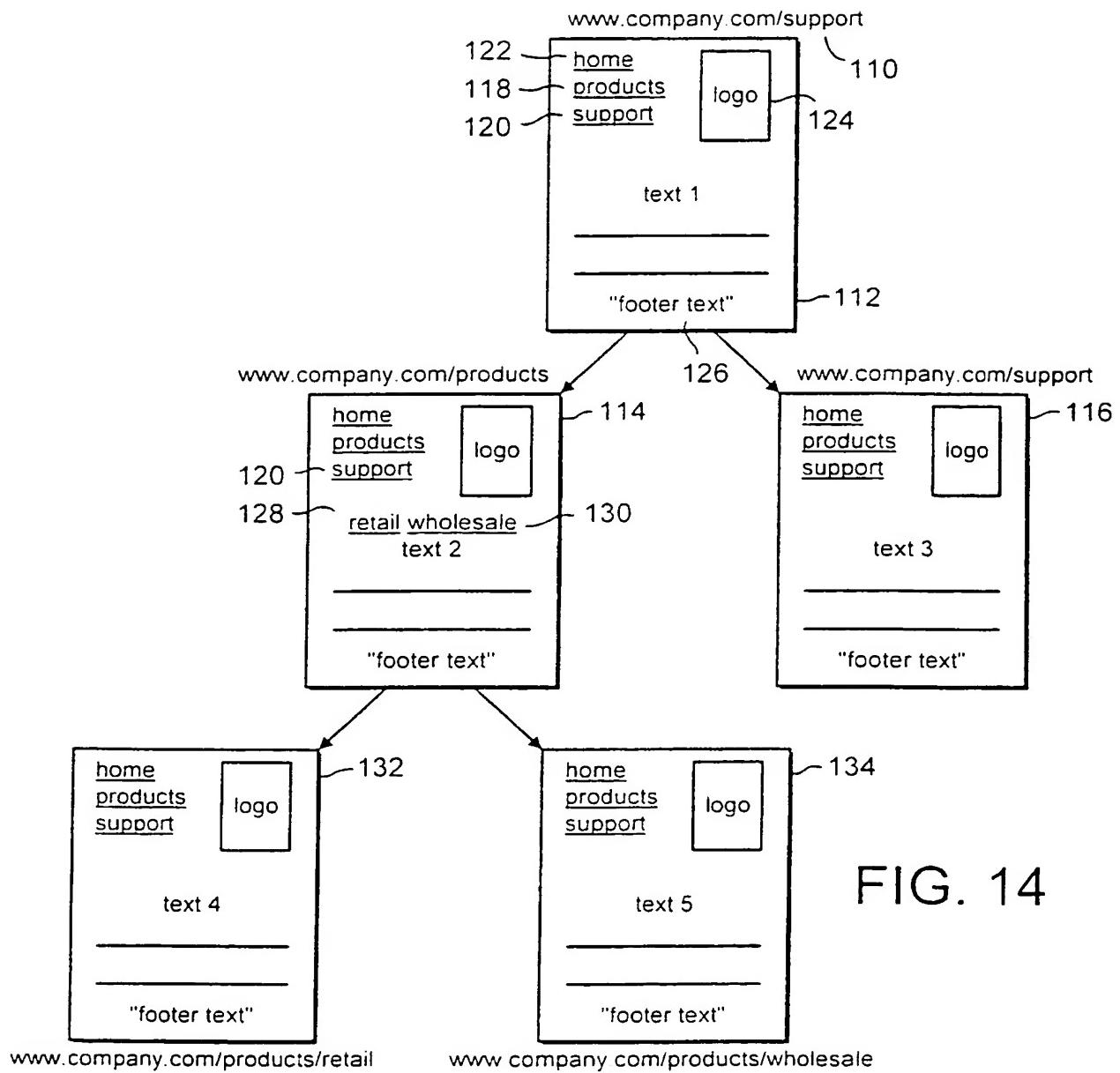


FIG. 14

SUBSTITUTE SHEET (RULE 26)

8 / 9

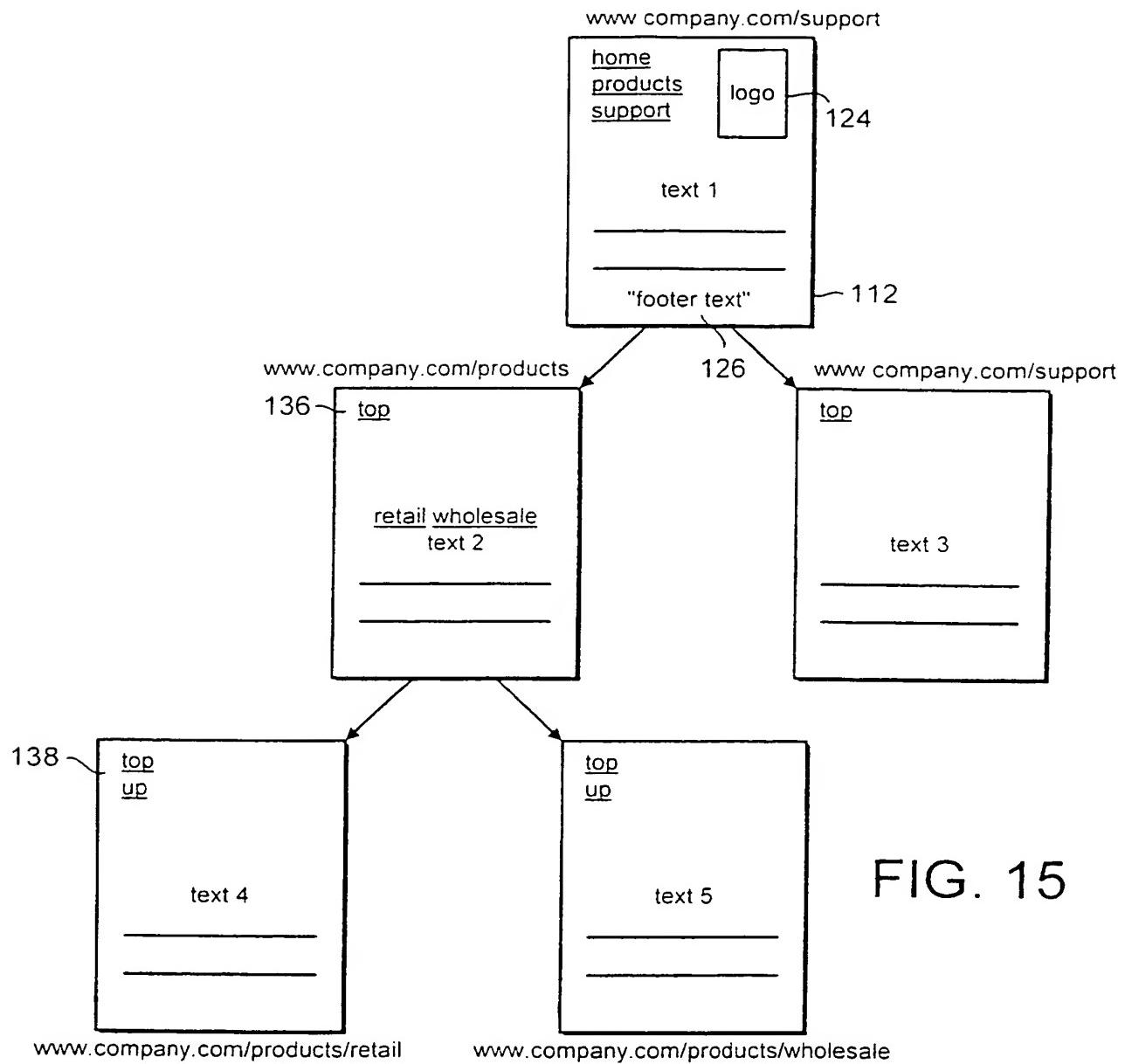


FIG. 15

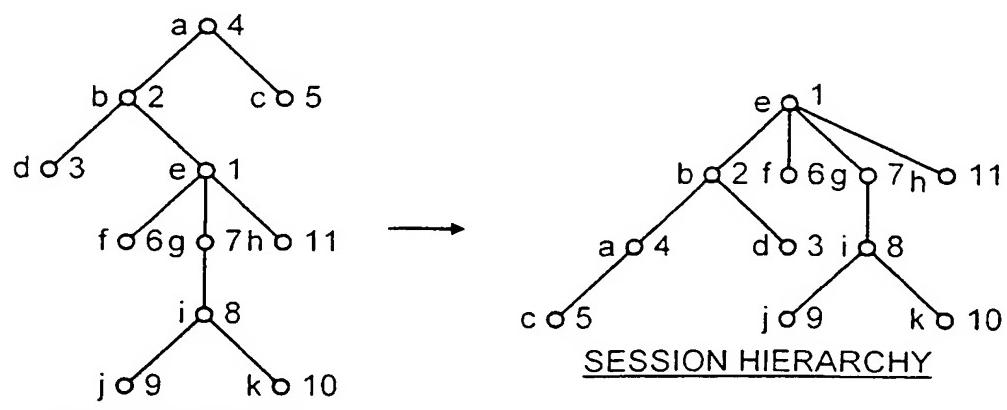


FIG. 16

SUBSTITUTE SHEET (RULE 26)

9 / 9

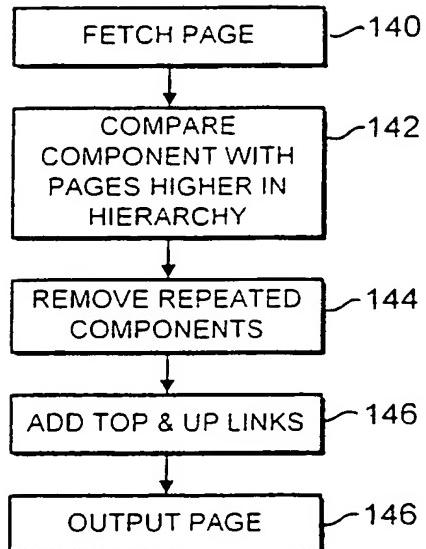


FIG. 17

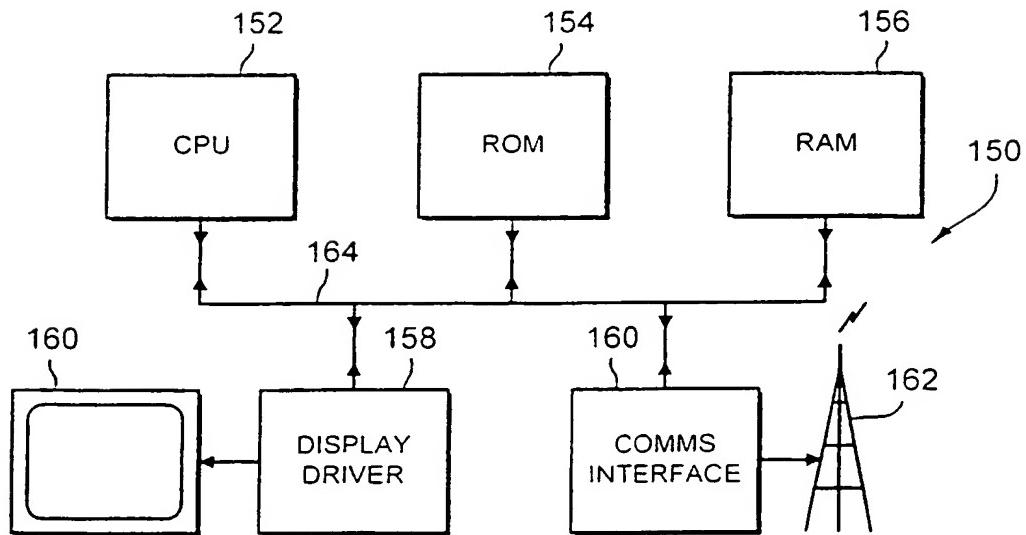


FIG. 18

SUBSTITUTE SHEET (RULE 26)

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
16 November 2000 (16.11.2000)

PCT

(10) International Publication Number
WO 00/68833 A3

(51) International Patent Classification⁷: **G06F 17/30**

(21) International Application Number: **PCT/GB00/01535**

(22) International Filing Date: 19 April 2000 (19.04.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

9910683.3	7 May 1999 (07.05.1999)	GB
9910684.1	7 May 1999 (07.05.1999)	GB
9910679.1	7 May 1999 (07.05.1999)	GB
9910682.5	7 May 1999 (07.05.1999)	GB
9910685.8	7 May 1999 (07.05.1999)	GB

[GB/GB]; 37 Bognor Road, Chichester, West Sussex PO19 2NG (GB). TRIBBECK, Jason, Paul [GB/GB]; 36 Netlecombe Avenue, Southsea, Portsmouth PO4 0QW (GB).

(74) Agent: **ROBINSON, Nigel, Alexander, Julian; D Young & Co., 21 New Fetter Lane, London EC4A 1DA (GB).**

(81) Designated States (*national*): JP, US.

(84) Designated States (*regional*): European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— with international search report

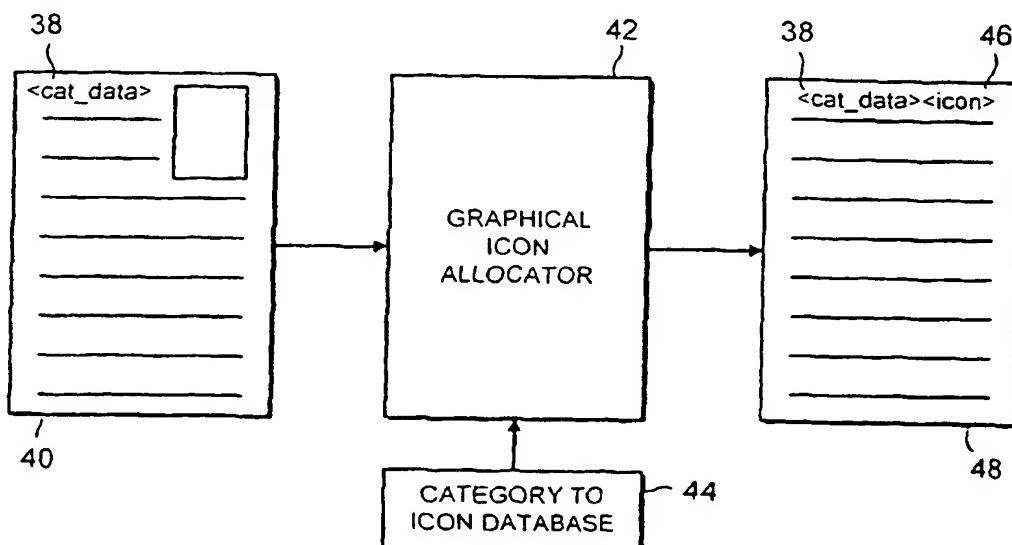
(71) Applicant (*for all designated States except US*): **ARGO INTERACTIVE LIMITED [GB/GB]**; 7 Dukes Court, Chichester, West Sussex PO19 2FX (GB).

(88) Date of publication of the international search report: 2 August 2001

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): **JELBERT, Richard**

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: CATEGORISING DATA

**WO 00/68833 A3**

(57) Abstract: A data processing system modifies a data file representing a document (14) and including link data items (24) by identifying keywords within the link data items and mapping these key words to category data (38) that is then inserted into the document. The document may be an internet web page and the link data item a hypertext link.

INTERNATIONAL SEARCH REPORT

Inte
onal Application No
PCT/GB 00/01535

A. CLASSIFICATION OF SUBJECT MATTER

IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, PAJ, IBM-TDB, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BICKMORE T W ET AL: "Digestor: device-independent access to the World Wide Web" COMPUTER NETWORKS AND ISDN SYSTEMS, NL, NORTH HOLLAND PUBLISHING. AMSTERDAM, vol. 29, no. 8-13, 1 September 1997 (1997-09-01), pages 1075-1082, XP004095305 ISSN: 0169-7552 page 1077, right-hand column, line 12 -page 1078, left-hand column, line 17 ---	1-26
A	US 5 727 159 A (KIKINIS DAN) 10 March 1998 (1998-03-10) column 7, line 13-43 column 10, line 9-35 ---	1-26 ---/---

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

* Special categories of cited documents :

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

& document member of the same patent family

Date of the actual completion of the international search

13 February 2001

Date of mailing of the international search report

21/02/2001

Name and mailing address of the ISA

European Patent Office, P B 5818 Patentlaan 2
 NL - 2280 HV Rijswijk
 Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.
 Fax. (+31-70) 340-3016

Authorized officer

Correia Martins, F

INTERNATIONAL SEARCH REPORT

Inte
onal Application No
PCT/GB 00/01535

C (Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No
A	EP 0 811 939 A (WEBTV NETWORKS INC) 10 December 1997 (1997-12-10) column 8, line 20-54 ---	1-26
A	JOHNSON D: "CONVERTING PC GUIS FOR NON PC DEVICES" CIRCUIT CELLUR INC, US, VERNON, CT, vol. 91, February 1998 (1998-02), pages 40-42, 44-45, XP000852859 ISSN: 0896-8985 page 42, left-hand column, line 1-30 -----	1-26

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inte
ntional Application No

PCT/GB 00/01535

Patent document cited in search report	Publication date	Patent family member(s)			Publication date
US 5727159	A 10-03-1998	CN 1218561	A	02-06-1999	
		EP 0892947	A	27-01-1999	
		JP 11508715	T	27-07-1999	
		WO 9738389	A	16-10-1997	
		US 6076109	A	13-06-2000	
<hr/>					
EP 0811939	A 10-12-1997	US 5918013	A	29-06-1999	
		AU 3375197	A	05-01-1998	
		JP 10228437	A	25-08-1998	
		WO 9746943	A	11-12-1997	
		US 6023268	A	08-02-2000	
		US 6141693	A	31-10-2000	
		US 5940074	A	17-08-1999	
		US 6073168	A	06-06-2000	
		US 5935207	A	10-08-1999	
		US 5996022	A	30-11-1999	
		US 5974461	A	26-10-1999	